L3: Entry 3 of 8 File: PGPB Mar 28, 2002

DOCUMENT-IDENTIFIER: US 20020038306 A1
TITLE: Method of managing slowly changing dimensions

Summary of Invention Paragraph:
[0002] The field of business applications of computer technology has seen many important changes over the last few years. With steadily growing computational power and data storage capacities of computer systems used for business data processing, the interest of the business community has shifted from transactional data management systems (on-line transaction processing systems, or OLTP systems, mostly supporting day-to-day business operations) and from relatively simple business data processing systems towards sophisticated business management systems, such as enterprise resource planning (ERP) systems, integrating at the enterprise level all facets and functions of the business, including planning, manufacturing, sales and marketing. An example of a business management software package of this scope is SAP R/3 System available from SAP AG (Germany) or its U.S. branch, SAP America, Inc.

Summary of Invention Paragraph:
[0003] Among various alternative approaches to business data management and analysis developed over the last few years, many are related to data warehousing. A data warehouse can be defined broadly as a subject-oriented collection of business data identified with a particular period of time (i.e., historically-oriented), as opposed to transactional (operational) databases dedicated to managing ongoing, day-to-day business activities. A scaled-down, usually single-subject oriented warehouse is sometimes referred to as a data mart. Data in a warehouse is normally gathered from a variety of sources (mostly various OLTP and legacy systems) and merged into a coherent whole. Data in a warehouse is usually stable, in that data is added to the warehouse but not removed. The latter feature, which is normally desirable to provide a more complete image of the business over time, may be absent from warehouses designed to keep data for a predetermined time span, with the oldest data being unloaded when the newest data is added.

Summary of Invention Paragraph:
[0004] As opposed to data stored in OLTP systems intended to support day-to-day operations and optimized mostly for the speed and reliability of transaction processing, data stored in a data warehouse or a data mart is intended to provide higher-level, aggregated views of the data, such as total sales by product line or region over a predetermined period of time, in support of business decision making. To provide consistently fast responses to such aggregate queries, data in a data warehouse or data mart must be structured in a manner facilitating the data synthesis, analysis, and consolidation.

Summary of Invention Paragraph:
[0008] A data warehouse or data mart is usually structured as a relational database, which can be seen as a collection of tables organized according to the dimensional model. Central to such a dimensionally-organized relational database (dimensional database) is a table known as the fact table, storing large amounts of aggregated business measures (facts), usually derived from transactional (operational) data of a business. Each row (record) of the fact table contains at least one aggregated business measure, for example total sales of a product during

a predetermined period of time, in addition to dimension keys identifying the product sold, time period during which the sales took place, geographic location of sales, and the like. In this example, characteristics like time, product and geographic location constitute business dimensions by which the data (facts) of the fact table are analyzed and the dimension keys of the fact record relate this record to relevant dimension tables. Additionally to the fact table, the dimensional database contains a number of dimension tables. A dimension table stores records of all members of a given dimension, each record (row of the dimension table) providing values of various attributes of members of the dimension, each attribute corresponding to a column of the dimension table. For example, for a client dimension, attributes may include client's key, name, address, telephone number, and the like. Examples of possible attributes of a product dimension are the product code, name, type, color, and size.

Summary of Invention Paragraph:
[0009] In the above model, each dimension table is related to the fact table by a single join (a star join schema), with dimensions considered to be independent. In real life applications, dimensions of a business dimensional model may not be and frequently are not independent. This is usually observed in dimensional models including a time dimension, when at least some of the remaining dimensions prove to be time-dependent, meaning that values of some attributes of certain members of such dimensions may change over time. For example, in a client dimension, addresses and/or telephone numbers of some clients may change occasionally. These changes are usually rare, meaning that a dimension undergoing such changes remains almost unchanged over time. Dimensions undergoing this kind of changes are known under the name of slowly changing dimensions. When the dimension tables of a data warehouse or data mart are updated with dimensional data extracted from transactional (operational) data, such changes are normally detected and have to be dealt with. Depending on how changes taking place in a given dimension over time are handled when updating its corresponding dimension table, three types of slowly changing dimensions, known as Type 1, Type 2, and Type 3, respectively, have been defined by Ralph Kimball and commonly accepted by the industry (see: Ralph Kimball, The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, John Wiley & Sons, Inc., New York 1996; Ralph Kimball et al., The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses, John Wiley & Sons, Inc., New York 1998).

Summary of Invention Paragraph:
[0010] The ability to deal with slowly changing dimensions is not always an integral part of software products known as ETL (Extract/Transform/Load) or ETD (Extract/Transform/Deliver) tools, which applications are used for constructing business data warehouses and data marts and for delivering transformed operational data into dimensional databases (data warehouses or data marts). The problem of slowly changing dimensions when delivering transformed data to a data mart was dealt with either manually or by writing an ad hoc piece of code particular to the star join schema at hand. DecisionStream, an ETL tool from Cognos BI suite, provides a new integrated method of dealing with slowly changing dimensions when building or updating a data mart, which method overcomes such prior art limitations.

Brief Description of Drawings Paragraph:
[0015] FIG. 1 is a screenshot showing the first panel of a dialog box for setting properties of a dimension template for a dimensional table according to a preferred embodiment of the invention;

Detail Description Paragraph:
[0018] For a given dimension, values of non-key attributes of certain members may be changing over time, without changing the value of the business key. For example, employees may change their department without changing their employee number (employee key), or the specification for a product may change without changing the

product code (product key). Such changes are mostly irrelevant to and may remain unnoticed in an operational system, which only contains data about the current state of the business at a given point in time. For example, a sales record of an operational system may show an office in which a sales representative worked at the time when the transaction was completed. This office may be different in a later sales record showing the same representative, if he moved in the meantime to another office. Such a change is recorded by the operational system in the explained manner, but has no or little meaning within the operational system. By contrast, a data warehouse is expected to hold data for a prolonged period of time and from the point of view of analysis of such data it may be important to know all the sales offices in which the sales representative has worked and when. This means that changes taking place over time in the sales representative dimension due, for example to a sales representative moving from one sales office to another, should be somehow recorded in the corresponding dimension table when this table is being updated with dimensional data showing such changes. In other words, in the context of a data warehouse, as opposed to an operational system, it is important to identify slowly changing dimensions (SCD) and to decide which historic values should be maintained. In this context, slowly changing dimensions became synonymous with the process and techniques for managing and preserving historic values for dimensions changing over time.

Detail Description Paragraph:
[0023] In the method according to the invention, the maintenance of both SCDs and surrogate keys is automated and does not require human intervention once an initial setup is completed. In a preferred embodiment, an internally assigned surrogate key is a 4-byte integer, meaning that more than 2 billion unique surrogate keys (positive integers) can be generated and assigned. By using an internally generated and assigned surrogate key, the uniqueness of the key can be ensured. Even if for a given dimension some externally assigned unique numerical key may exist, such as a social security number, that number may be missing or incorrect when the data is entered into the system. An internally assigned surrogate key always exists and is guaranteed unique.

Detail Description Paragraph:
[0024] Even though operational databases from which data are extracted, transformed and delivered to a data mart may sometimes use surrogate keys (e.g., employee number) which can be passed into the data mart, these operational surrogate keys normally cannot and should not be used as data mart surrogate keys. For example, when merging entities from separate operational systems, each with its own operational surrogate key, it may be preferable to assign a single surrogate key to the merged entity, e.g., to uniquely identify a single customer originally identified by its checking account, savings account and insurance policy numbers, each of them being a unique operational surrogate keys. In such a case the operational surrogate keys, when transferred into the data mart, may still play the role of natural (business) keys and can be used for queries. On the other hand, a single member in a data mart, for example an employee or a product, may have several data mart surrogate keys assigned over time to deal with slowly changing dimensions.

Detail Description Paragraph:
[0032] In a preferred embodiment, the method of managing slowly changing dimensions according to the invention is embedded in an ETL application running under an operating system, preferably under the MS Windows operating system, using facilities and methodologies of the Windows environment well known to those skilled in the art, such as the point-and-click graphical user interface, as well as standard input and output devices, such as a mouse and a keyboard. In this environment, a dimension template assigning a behavior type to each column of a dimension table is created using a suitable dialog box, as shown in FIG. 1. This dialog box contains two panels, associated with "General" and "Attributes" tabs. In FIG. 1 the "General" panel is in the foreground, brought into this position by

clicking at the corresponding tab. This panel contains three text fields into which general information about the template can be entered: name (name of the template, which is mandatory and may or may not be the same as the name of the dimension table associated with the template), business name (name of the business to which the dimension table pertains), and description (for description of the dimension table, template, business, etc.). The last two fields are optional.

Detail Description Paragraph:
[0034] In the following, the method of the present invention will be further explained for a simple dimensional database based on a simple star join schema, i.e., consisting of a single fact table and a number of dimension tables related to the fact table by a single join, each dimension table corresponding to a business dimension. The basic requirements for the dimension table is that it has a column for a business key, may have columns which are used to maintain the table, and may have further columns representing various attributes of the key.

Previous Doc Next Doc Go to Doc#